TF-IDF 法に基づくタグの自動付与

河上 哲也 千種康民† 服部泰造‡
東京工科大学 コンピュータサイエンス学部† 東京国際大学‡

1. 研究の背景と目的

現在, 世の中には多種多様なドキュメントがあるが, それらを分類する方法はひとつの課題である。最近では, サイト運営者が分類（カテゴリ分け）するのではなく, 利用者がコンテンツにタグをつけて分類する方法が登場した.

フォークソノミは, コンテンツに対して 人がタグ（キーワード）を多数付加していくことで利用者本位な分類をしていく手法である. 通常フォークソノミは人の手により行われるため, 大変な労力がかかる.

そこで, 本研究ではフォークソノミにおけるタグ付けを自動化する研究を行った.

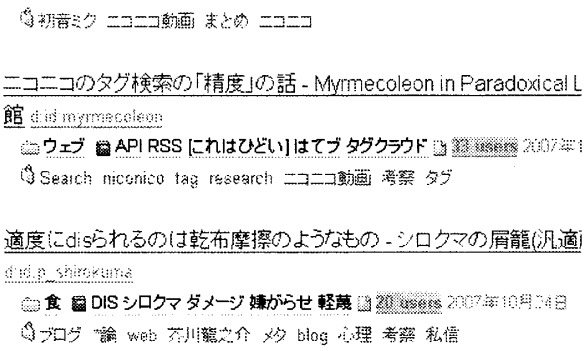

図 1フォークソノミの例

2. 特徴

フォークソノミによる分類は, カテゴリ分けにとらわれることのない自由な分類が可能となる. タグの綴りの間違えや, 同義語といった問題点を解決することができる. 意味のないタグ付がなくなる. タグ付が自動化されるため, 労力が軽減される. タグ付がされていない既存のテキストに適応できる. といった特徴がある.

Automatic Ttagging Based on TF-IDF
Tetsuya KAWAKAMI †, Yasutami CHIGUSA † and
Taizoh HATTORI ‡
† Tokyo University of Technorogy
‡Tokyo International University

3. 実装方法

対象ドキュメントからキーワードを抽出し, それを評価してタグを出力するシステムを実装する.

キーワードを抽出するためにドキュメントを形態素解析を用いて形態素に分割し, それらの単語を TF-IDF 法により重みづけをして評価を行う. 最終的にスコアの高い単語をタグとして出力する

TF-IDF 法は単語の出現頻度（tf）と単語の希少度（idf）で単語を評価する手法である. スコア（w）の高い単語をそのドキュメントのタグとして出力する.

$$tfidf(d,t) = tf(d,t) \cdot idf(t) \quad (1)$$

$$idf(t) = log(\frac{N}{df(t)}) \quad (2)$$

d:入力文書 t:対象となる単語 tf:単語の出現頻度 df:単語の文書出現頻度 N:総文書数

評価したタグはデータベースにしておき、同じタグが付けられたときや, タグによる検索が行われた時にタグを再評価する. これにより使用頻度の高いタグは評価が高くなり, 使用頻度の低いタグは評価が下がる.

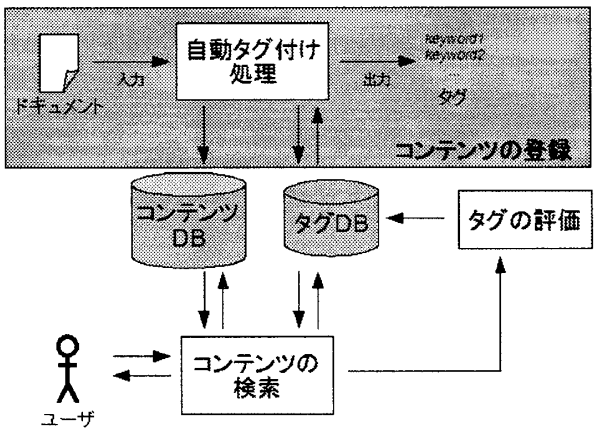

図 2 システム構成図

4. タグのスコア付けの結果

図 3-6 は§3 の手法により求めたスコアをもとに，入力文書のテキストの各単語のフォントサイズ変化をさせることで，もとの文書を視覚的にわかりやすくした例である．また、結果は品詞ごとに色を変えてある．

5. まとめ

ドキュメントの分類手法の一つであるフォークソノミをとりあげ、その問題点及びその解決方法として自動タグ付による方法を提案し、その手法についてとりあげた．また，スコアづけの評価を行うための視覚化についてとりあげた．

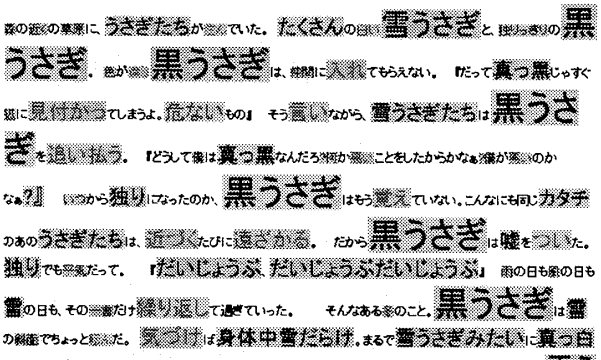

図 3 童話を視覚化した例

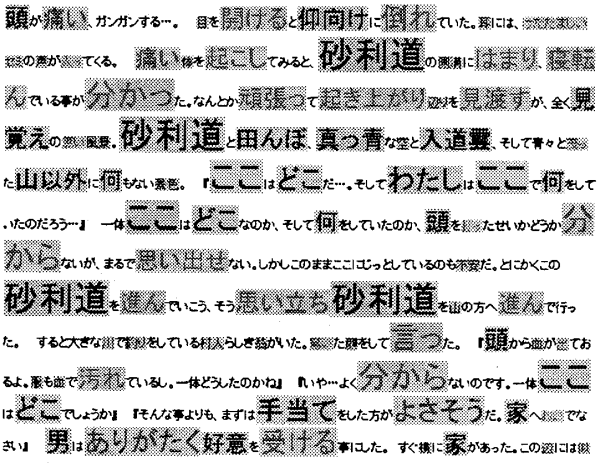

図 4 ショートストーリーを視覚化した例

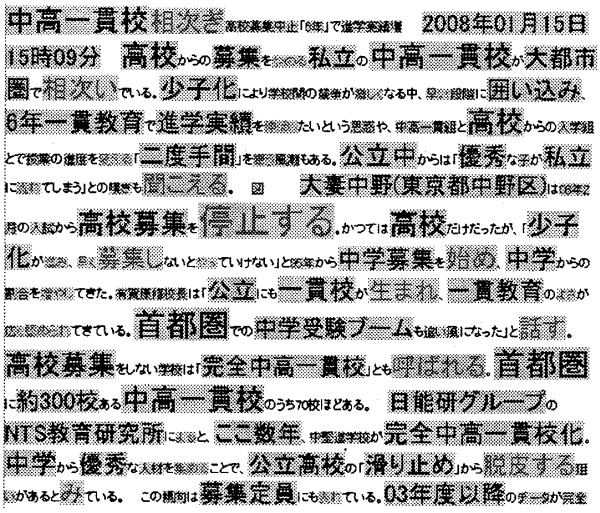

図 5 ニュースを視覚化した例

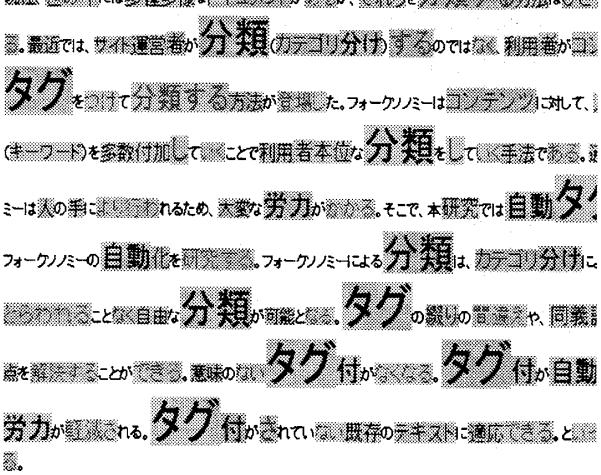

図 6 本稿の 2 節を視覚化した例

参考文献

[1] 西村 悟, 千野 晋平, 三木 光範, 廣安 知之:【IT 用語】Folksonomy ～タグで繋がるみんなの分類～:
http://mikilab.doshisha.ac.jp/dia/research/report/2005/0911/001/report20050911001.html :2005/10/24

[2] taku-ku:MeCab Yet Another Part-of-Speech and Morphological Analyzer:
http://mecab.sourceforge.net/

[3]Yahoo:Yahoo!デベロッパーネットワーク:
http://developer.yahoo.co.jp/